Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria

I. King Jordan, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

The "knockout-rate" prediction holds that essential genes should be more evolutionarily conserved than are nonessential genes. This is because negative (purifying) selection acting on essential genes is expected to be more stringent than that for nonessential genes, which are more functionally dispensable and/or redundant. However, a recent survey of evolutionary distances between *Saccharomyces cerevisiae* and *Caenorhabditis elegans* proteins did not reveal any difference between the rates of evolution for essential and nonessential genes. An analysis of mouse and rat orthologous genes also found that essential and nonessential genes evolved at similar rates when genes thought to evolve under directional selection were excluded from the analysis. In the present study, we combine genomic sequence data with experimental knockout data to compare the rates of evolution and the levels of selection for essential versus nonessential bacterial genes. In contrast to the results obtained for eukaryotic genes, essential bacterial genes appear to be more conserved than are nonessential genes over both relatively short (microevolutionary) and longer (macroevolutionary) time scales.

Rates of evolution vary tremendously among protein-coding genes. Molecular evolutionary studies have revealed an ~1000-fold range of nonsynonymous substitution rates (Li and Graur 1991). The strength of negative (purifying) selection is thought to be the most important factor in determining the rate of evolution for the protein-coding regions of a gene (Kimura 1983; Ohta 1992; Li 1997). Consistent with this idea, Alan Wilson and colleagues (1997) proposed that essential genes should evolve more slowly than nonessential genes. This is the so-called "knockout-rate" prediction (Hurst and Smith 1999). "Essential" and "nonessential" are classic molecular genetic designations that relate to the functional significance of a gene with respect to its effect on organismic fitness. A gene is considered to be essential if a knock-out results in (conditional) lethality or infertility. On the other hand, nonessential genes are those for which knock-outs yield viable and fertile individuals. It was reasoned that purifying selection should be more intense for essential genes because they are, by definition, less functionally dispensable and/or redundant than are nonessential genes. Given the role of purifying selection in determining evolutionary rates, the greater levels of purifying selection on essential genes should be manifest as a lower rate of evolution relative to that of nonessential genes.

To systematically evaluate the relationship between the fitness effects of genes and their rates of evolution, a combination of a substantial amount of experimental knock-out data and sequence data from numerous genes is required. Only recently has enough data accumulated to allow for tests of the straightforward and seemingly intuitive knock-out rate prediction. However, examinations of sequence data with respect to this prediction have yielded equivocal results. For example, a survey of substitution rates for mouse and rat orthologous genes appeared to indicate a slower rate of evolution for essential genes. But when genes thought to evolve under directional selection were excluded from the analysis, essential and nonessential genes were found to evolve at similar rates (Hurst and Smith 1999). A more recent analysis of the evolutionary distances between *Saccharomyces cerevisiae* and *Caenorhabditis elegans* proteins did indicate that the fitness effect of a protein influences its rate of evolution (Hirsh and Fraser 2001). Nevertheless, this study (Hirsh and Fraser 2001) was also unable to reveal any difference between the rates of evolution for essential and nonessential genes.

The results from both of these studies were taken to indicate that the fitness differences between essential and nonessential genes do not influence evolutionary rates to the extent that was expected. However, the studies relied on the analyses of relatively few genes (n = 175 and n = 287, respectively) and comparisons between species that diverged at least tens of millions of years ago. It might be the case that these results reflect a lack of power and sensitivity of the approaches that were used. The recent availability of complete genome sequences from different strains of the same bacterial species provides an opportunity to address the issue with an unprecedented level of resolution. In the present study, to test the knockout-rate prediction, the relationship between the fitness class of genes (essential versus nonessential) and their rate of evolution was assessed for three bacterial species: *Escherichia coli*, *Helicobacter pylori*, and *Neisseria meningitidis*, for each of which at lease two complete genome sequences are available.

RESULTS AND DISCUSSION

The Profiling of the *E. coli* Genome (PEC) database (http://www.shigen.nig.ac.jp/ecoli/pec/) was used to characterize *E. coli* genes as essential, nonessential, or undetermined. *E. coli* genes in this database are characterized as essential or nonessential based largely on experimental (null mutations) evi-

[1]Corresponding author.
E-MAIL koonin@ncbi.nlm.nih.gov; FAX (301) 435-7794.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.87702. Article published online before print in May 2002.

12:962–968 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

dence. In addition to the use of experimental evidence, a much smaller number of genes were designated as essential or nonessential based simply on their known function (see Methods). For example, the genes for ribosomal proteins were assumed to be essential, whereas genes involved in flagellation, motility, and chemotaxis were classified as nonessential. Comparisons between essential and nonessential *E. coli* genes were performed using all of the data available from the PEC database and with a reduced data set that contained only essential and nonessential genes for which experimental evidence existed. Sets of orthologous protein sequences and the corresponding nucleotide sequences shared by the two completely sequenced *E. coli* strains were identified and aligned, and nucleotide sequence alignments were used to calculate the synonymous (Ks) and nonsynonymous (Ka) substitution rates (see Methods). Values of Ks and Ka were compared for *E. coli* genes designated as either essential or nonessential. Analysis of these genes showed that the average Ks and Ka were significantly lower for essential genes than for nonessential genes (Table 1). This is the case for both the entire data set and the reduced set containing only experimentally characterized genes (Table 1). The reduction in Ka for essential genes indicates a reduction in the intraspecific rate of essential protein evolution, and the reduction in Ks is consistent with the positive correlation between Ks and Ka (Table 1). Such a correlation between Ks and Ka values has also been observed for several other species (Wolfe and Sharp 1993; Ohta and Ina 1995; Makalowski and Boguski 1998). This correlation could reflect a mechanistic bias in mutation or indicate that synonymous sites are also subject to some degree of selection (or both). However, the significantly lower value of Ka/Ks for essential genes (Table 1) indicates that the difference between the rates of evolution for the two gene classes is more pronounced for Ka than for Ks and is consistent with more stringent negative selection against amino acid replacements acting on essential genes. The undetermined genes, which were not included in either the essential or the nonessential class, had somewhat greater average Ka and Ka/Ks values than those of the nonessential genes (Table 1).

It is a formal possibility that a relatively few genes with extreme values contributed disproportionately to the average Ka, Ks, and Ka/Ks and were thus largely responsible for the difference between these average values for essential and nonessential *E. coli* genes. The fact that we observed the significant difference between the essential and nonessential genes after the nonexperimentally assigned genes for ribosomal proteins were removed from the essential set (Table 1) argues against this possibility. Indeed, although the ribosomal protein genes are highly conserved and had substantially lower average values of Ka, Ks, and Ka/Ks than those for the rest of the essential genes (data not shown), these genes alone clearly did not account for the difference between the essential and nonessential set.

To further assess the effect of potential biases in the essential gene set on the observed differences in evolutionary rates, the values of Ka, Ks, and Ka/Ks for essential and nonessential *E. coli* genes were reanalyzed with a bootstrap test (Methods). The frequency distributions of bootstrapped average Ka, Ks, and Ka/Ks show clear distinctions between essential and nonessential genes (Fig. 1). In addition, for each bootstrap replicate, the average values of essential and nonessential genes were calculated, and the significance of the difference between the essential and nonessential average values was assessed. For Ka and Ks each, all 1000 replicates differed at the $P < 0.01$ level; for Ka/Ks, 968 replicates differed at the $P < 0.01$ level. Thus, the results of the bootstrap analysis reject the possibility that the average values of Ka, Ks, and Ka/Ks for essential and/or nonessential genes are greatly influenced by a few genes with extreme values.

Evolutionary conservation over longer time scales was measured by calculating the phyletic distribution parameter of essential

Table 1. The Rates of Synonymous (Ka) and Nonsynonymous (Ks) Nucleotide Substitutions for Essential Versus Nonessential Bacterial Genes

	Ks (±se)[a]	Ka (±se)[b]	r[c]	Ka/Ks
Escherichia coli—all[d]	26.99×10^{-3}	1.11×10^{-3}	0.35	4.50×10^{-2}
Essential (n = 205)	($\pm 2.1 \times 10^{-3}$)	($\pm 0.1 \times 10^{-3}$)		($\pm 0.7 \times 10^{-2}$)
Nonessential (n = 1794)	51.0×10^{-3}	3.60×10^{-3}	0.44	8.40×10^{-2}
	($\pm 1.0 \times 10^{-3}$)	($\pm 0.2 \times 10^{-3}$)		($\pm 0.2 \times 10^{-2}$)
Undetermined (n = 1107)	48.6×10^{-3}	5.00×10^{-3}	0.47	11.67×10^{-2}
	($\pm 1.3 \times 10^{-3}$)	($\pm 0.3 \times 10^{-3}$)		($\pm 0.5 \times 10^{-2}$)
Significance of the difference[f]	$P = 0$	$P = 5.6 \times 10^{-16}$	na	$P = 6.4 \times 10^{-6}$
E. coli—experimental[e]	35.33×10^{-3}	1.36×10^{-3}	0.33	4.73×10^{-6}
Essential (n = 150)	($\pm 2.5 \times 10^{-3}$)	($\pm 0.2 \times 10^{-3}$)		($\pm 0.7 \times 10^{-2}$)
Nonessential (n = 1736)	51.29×10^{-3}	3.60×10^{-3}	0.44	8.27×10^{-2}
	($\pm 1.0 \times 10^{-3}$)	($\pm 0.2 \times 10^{-3}$)		($\pm 0.4 \times 10^{-2}$)
Significance of the difference[f]	$P = 1.2 \times 10^{-7}$	$P = 1.2 \times 10^{-8}$	na	$P = 5.6 \times 10^{-4}$
Helicobacter pylori	111.33×10^{-3}	12.89×10^{-3}	0.43	11.32×10^{-2}
Essential (n = 98)	($\pm 4.1 \times 10^{-3}$)	($\pm 1.1 \times 10^{-3}$)		($\pm 0.9 \times 10^{-2}$)
Nonessential (n = 130)	135.24×10^{-3}	21.64×10^{-3}	0.30	16.14×10^{-2}
	($\pm 3.1 \times 10^{-2}$)	($\pm 1.4 \times 10^{-3}$)		($\pm 0.8 \times 10^{-2}$)
Significance of the difference[f]	$P = 3.5 \times 10^{-6}$	$P = 6.1 \times 10^{-7}$	na	$P = 6.1 \times 10^{-4}$
Neisseria meningitidis	65.37×10^{-3}	4.76×10^{-3}	0.68	7.32×10^{-2}
Essential (n = 98)	($\pm 6.9 \times 10^{-3}$)	($\pm 0.7 \times 10^{-3}$)		($\pm 1.0 \times 10^{-2}$)
Nonessentials (n = 130)	91.56×10^{-3}	9.60×10^{-3}	0.49	17.65×10^{-2}
	($\pm 5.9 \times 10^{-3}$)	($\pm 0.9 \times 10^{-3}$		($\pm 1.9 \times 10^{-2}$)
Significance of the difference[f]	$P = 1.6 \times 10^{-4}$	$P = 1.0 \times 10^{-7}$	na	$P = 4.2 \times 10^{-5}$

[a]The average synonymous substitution (nucleotide substitutions that do not change the encoded amino acid sequence) rate (Ks) for all orthologous genes within a given fitness class and species is shown with the standard error (in parentheses).
[b]The average nonsynonymous substitution (nucleotide substitutions that change the encoded amino acid sequence) rate (Ka) for all orthologous genes within a given fitness class and species is shown with the standard error (in parentheses).
[c]Correlation coefficient between the Ks and Ka values.
[d]This set includes all of the *E. coli* genes characterized as essential or nonessential (on the basis of experimental and functional information).
[e]This set includes only the *E. coli* genes characterized as essential or nonessential on the basis of experimental evidence.
[f]Statistical significance of the difference between the essential and nonessential classes for a given measurement as determined using the Mann-Whitney *U* test.

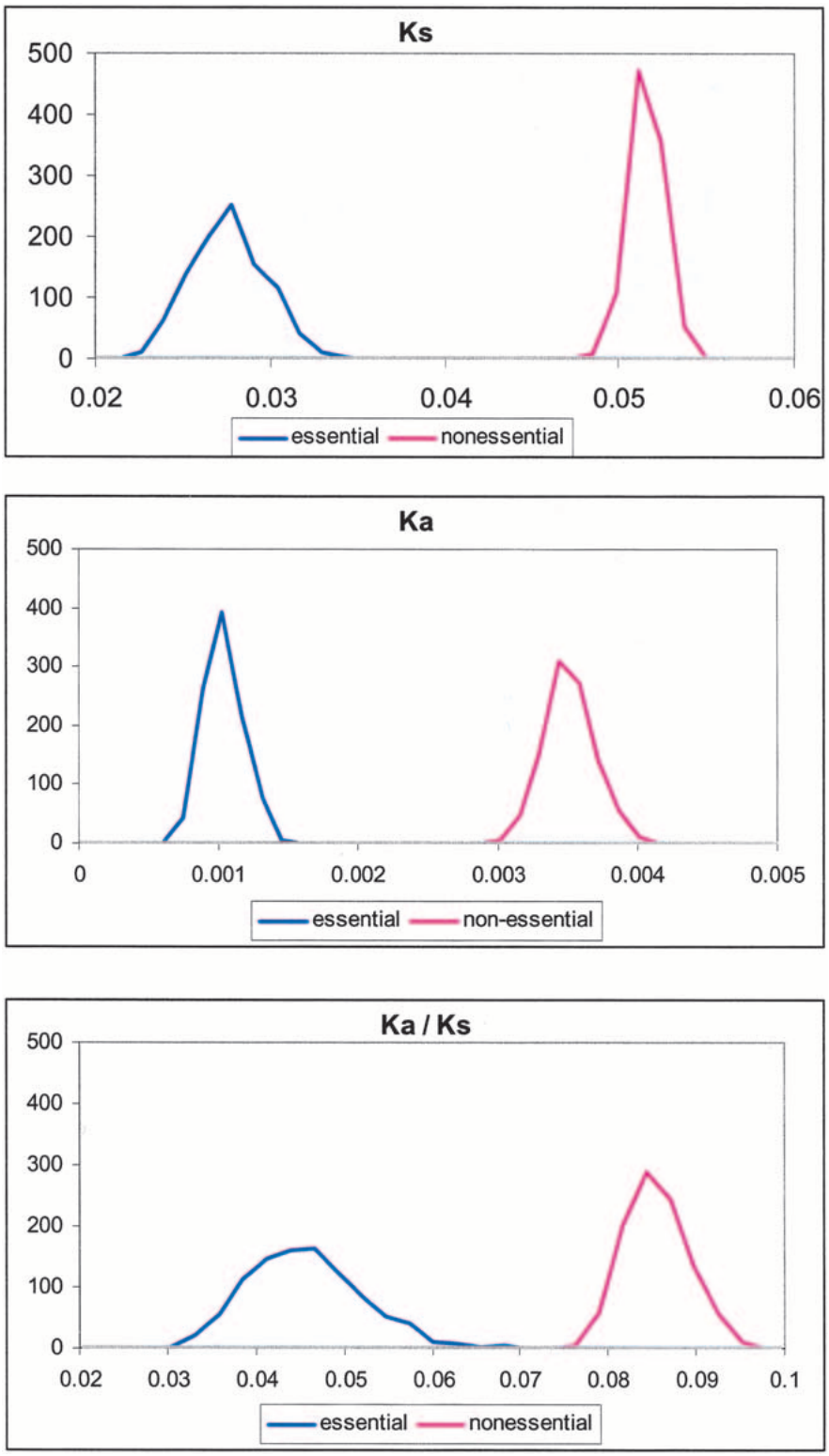

Figure 1 Bootstrap test of the average rates of synonymous (Ks), nonsynonymoys (Ka), and Ka/Ks for essential and nonessential *Escherichia coli* genes. Frequency distributions for the average values of Ka, Ks, and Ka/Ks for 1000 bootstrap replicates are shown.

versus nonessential *E. coli* genes. This parameter indicates the extent to which orthologs of a gene are distributed among the 26 taxonomic groups in the Clusters of Orthologous Groups (COGs) database (see Methods). Orthologs of essential *E. coli* genes are more broadly distributed ($P < 10^{-10}$, Mann-Whitney *U* test) among bacterial and archaeal species than are orthologs of nonessential *E. coli* genes (Fig. 2).

Orthologs of essential and nonessential genes from *E. coli* (all genes in each category, without removing nonexperimentally characterized genes) were identified in *H. pylori* and *N. meningitidis* (see Methods). The classification of these orthologs as essential or nonessential in *E. coli* was taken as an approximation of their classification in *H. pylori* and *N. meningitidis*, and the same evolutionary comparisons were performed on them. For both species, the rates of Ks and Ka for 98 predicted essential genes were significantly lower than the rates for 130 predicted nonessential genes (Table 1). The differences among Ks and Ka values between all three species surveyed merely reflect the fact that for each species, the time to common ancestry for the two sequenced genomes, in all likelihood, differs significantly. Also like in the case of *E. coli*, the average Ka/Ks values are lower for the predicted essential genes (Table 1). Orthologs of predicted essential genes in these species were also more broadly phyletically distributed ($P = 1.6 \times 10^{-9}$ Mann-Whitney *U* test) than are orthologs of predicted nonessential genes (Fig. 2). An additional aspect of these comparisons was that *N. meningitidis* and particularly *H. pylori* had significantly greater values of Ks, Ka, and Ka/Ks than those for *E. coli*, such that, for example, even essential genes of *H. pylori* appeared to evolve faster than nonessential genes of *E. coli* (Table 1). A detailed examination of these differences is beyond the scope of the present work, but it seems interesting to speculate that they might reflect the parasitic lifestyle of *N. meningitidis* and *H. pylori*.

Finally, sets of interspecific (*E. coli*, *H. pylori*, and *N. meningitidis*) orthologous proteins were aligned, and the average pair-wise evolutionary distance for each set was calculated (see Methods). Essential proteins show a significantly lower

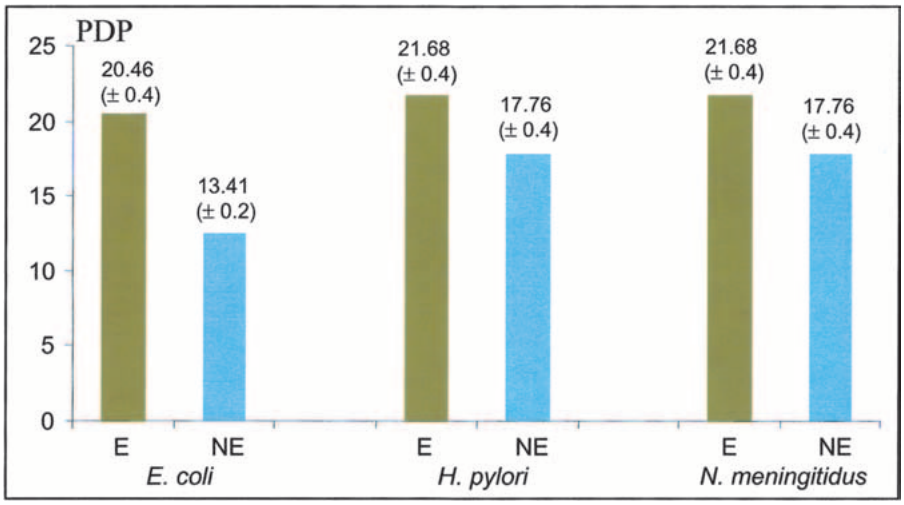

Figure 2 Average phyletic distribution parameter values for essential (E) and nonessential (NE) genes in the three bacterial species surveyed. Average values (±SE) are shown above the bars. For the definition of the phyletic distribution parameter (PDP), see Methods.

($P = 6.3 \times 10^{-6}$ Mann-Whitney U test) average level of per site amino acid sequence variation (avgerage ± SE = 4.85 ± 0.25) among these three species than do nonessential genes (average ±SE = 6.83 ± 0.34).

A previous comparison of the rates of evolution for essential versus nonessential genes in mammals initially revealed significantly lower rates of evolution for essential genes (Hurst and Smith 1999). However, when genes involved in the immune system, which are thought to evolve under diversifying (positive) selection, were removed from the analyzed data set, this difference disappeared. This result was attributed to substantial differences in the rates of evolution for different functional classes of genes. We sought to explore the potential contribution of this phenomenon to the observed difference between evolutionary rates of essential versus nonessential bacterial genes by breaking down the *E. coli* genes into four broad functional categories (see Methods): (1) information processing and storage, (2) cellular processes, (3) metabolism, and (4) poorly characterized. Not unexpectedly, comparison of the numbers of genes of each functional class that were designated as essential or nonessential revealed a nonrandom distribution (data not shown). For example, information storage and processing genes are vastly overrepresented among essential genes, whereas there are far fewer poorly characterized genes than would be expected by chance in this same set. Notably, however, there were no significant differences in the evolutionary rates among information processing, cellular processes, and metabolic categories within the essential, nonessential, or undetermined sets; only the poorly characterized genes appeared to evolve significantly faster (Table 2). In three of the four functional classes, the essential genes had significantly lower values of Ka and tended to have significantly lower values of Ks and Ka/Ks than did the nonessential genes. In addition, in each of the classes, the undetermined genes were found to evolve even faster than nonessential ones, and this difference was significant on two occasions (Table 2). The only exception to this pattern was among the poorly characterized genes. Poorly characterized essential genes do have substantially lower rates of evolution than do nonessential genes, but the low number of these essential genes (n = 12) results in a statistical comparison that lacks power. Thus, the slower rate of evolution of essential genes compared with nonessential genes in bacteria appears to be a general phenomenon that is not limited to a particular functional category of genes.

Despite intense scrutiny of the factors that influence the rate of protein evolution, confirmation of the straightforward prediction that essential genes should be more evolutionarily conserved than are nonessential genes (Wilson et al. 1977) has proven elusive. A recent study of protein variation between *S. cerevisiae* and *C. elegans* did reveal a significant linear relationship between the protein's level of dispensability (fitness class as determined in *S. cerevisiae*) and their rate of evolution (Hirsh and Fraser 2001). This is consistent with the idea that the rate of the evolution of a gene depends on its contribution to the fitness of the organism. However, this same study did not find any difference between the rates of evolution of essential versus nonessential genes. This lack of difference was attributed to the fact that ablation of many of the nonessential genes may have enough of an effect on organismal fitness to render them evolutionarily equivalent to essential genes. However, in the present study, a dense sampling of orthologous genes, facilitated by the completion of multiple bacterial genome sequences, allowed us to show that essential genes in bacteria are more conserved than are nonessential genes. It is unclear whether the difference between the findings for eukaryotic and bacterial genes is merely an effect of the sampling or a reflection of a real distinction between the evolutionary modes of these two domains of life.

A similar survey of mouse and rat orthologous genes initially found a difference between the evolutionary rates of essential and nonessential genes, but when genes thought to evolve under directional selection were excluded from the analysis, this distinction disappeared (Hurst and Smith 1999). This raises the question of whether the differences between essential and nonessential genes reported here is because of positive selection being more prevalent in nonessential genes or to purifying selection being more stringent among essential genes. We did not find any evidence of positive selection (i.e., Ka/Ks > 1) in our comparisons. However, Ka/Ks > 1 is an extremely conservative criterion, which will reveal only cases of strong positive selection acting on large portions of genes. Therefore, it remains a formal possibility that the differences described here could be owing in small part to differences between essential and nonessential genes in the frequency of positive selection. However, purifying selection is clearly the rule in protein evolution, as evidenced by the Ka/Ks values reported here and in numerous other studies. Thus, consistent with the knockout-rate prediction (Wilson et al. 1977), differences in levels of purifying selection have certainly had the decisive role in determining the different rates of evolution of essential and nonessential bacterial genes.

Until this time, attempts to verify the knockout-rate pre-

Table 2. The Rates of Synonymous (Ks) and Nonsynonymous (Ka) Nucleotide Substitutions among Different Functional Categories for Essential, Nonessential, and Undetermined *Escherichia coli* Genes

	Ks (±se)[a]	Ka (±se)[b]	Ka/Ks (±se)
Information storage and processing			
Essential (n = 113)	19.42×10^{-3} ($\pm 2.3 \times 10^{-3}$)	0.96×10^{-3} ($\pm 0.2 \times 10^{-3}$)	3.89×10^{-2} ($\pm 0.8 \times 10^{-2}$)
Nonessential (n = 222)	48.16×10^{-3} ($\pm 2.8 \times 10^{-3}$)	3.50×10^{-3} ($\pm 0.4 \times 10^{-3}$)	8.72×10^{-2} ($\pm 1.1 \times 10^{-2}$)
Undetermined (n = 113)	52.31×10^{-3} ($\pm 4.4 \times 10^{-3}$)	4.62×10^{-3} ($\pm 0.6 \times 10^{-3}$)	11.69×10^{-2} ($\pm 1.8 \times 10^{-2}$)
Significance of the difference[c] essential vs. nonessential	$P = 1.6 \times 10^{-15}$	$P = 1.6 \times 10^{-7}$	$P = 6.1 \times 10^{-3}$
Significance of the difference[c] essential vs. undetermined	$P = 3.2 \times 10^{-12}$	$P = 1.3 \times 10^{-10}$	$P = 1.3 \times 10^{-4}$
Significance of the difference[c] nonessential vs. undetermined	NS	NS	NS
Cellular processes			
Essential (n = 46)	37.57×10^{-3} ($\pm 3.5 \times 10^{-3}$)	1.32×10^{-3} ($\pm 0.3 \times 10^{-3}$)	4.72×10^{-2} ($\pm 1.5 \times 10^{-2}$)
Nonessential (n = 408)	47.17×10^{-3} ($\pm 1.9 \times 10^{-3}$)	3.01×10^{-3} ($\pm 0.3 \times 10^{-3}$)	7.86×10^{-2} ($\pm 0.7 \times 10^{-2}$)
Undetermined (n = 154)	40.92×10^{-3} ($\pm 2.1 \times 10^{-3}$)	4.10×10^{-3} ($\pm 0.5 \times 10^{-3}$)	8.80×10^{-2} ($\pm 0.9 \times 10^{-2}$)
Significance of the difference[c] essential vs. nonessential	NS	$P = 3.6 \times 10^{-2}$	NS
Significance of the difference[c] essential vs. undetermined	NS	$P = 2.0 \times 10^{-4}$	$P = 3.7 \times 10^{-4}$
Significance of the difference[c] nonessential vs. undetermined	NS	$P = 2.8 \times 10^{-3}$	$P = 3.2 \times 10^{-4}$
Metabolism			
Essential (n = 34)	37.18×10^{-3} ($\pm 8.1 \times 10^{-3}$)	1.09×10^{-3} ($\pm 0.3 \times 10^{-3}$)	3.19×10^{-2} ($\pm 1.1 \times 10^{-2}$)
Nonessential (n = 725)	54.76×10^{-3} ($\pm 1.5 \times 10^{-3}$)	2.52×10^{-3} ($\pm 0.1 \times 10^{-3}$)	5.92×10^{-2} ($\pm 0.4 \times 10^{-2}$)
Undetermined (n = 241)	59.32×10^{-3} ($\pm 3.2 \times 10^{-3}$)	4.50×10^{-3} ($\pm 0.3 \times 10^{-3}$)	9.80×10^{-2} ($\pm 0.8 \times 10^{-2}$)
Significance of the difference[c] essential vs. nonessential	$P = 1.2 \times 10^{-4}$	$P = 1.2 \times 10^{-3}$	$P = 2.7 \times 10^{-3}$
Significance of the difference[c] essential vs. undetermined	$P = 8.2 \times 10^{-5}$	$P = 4.9 \times 10^{-8}$	$P = 1.7 \times 10^{-6}$
Significance of the difference[c] nonessential vs. undetermined	NS	$P = 3.6 \times 10^{-13}$	$P = 8.2 \times 10^{-11}$
Poorly Characterized			
Essential (n = 12)	28.75×10^{-3} ($\pm 6.0 \times 10^{-3}$)	1.75×10^{-3} ($\pm 0.3 \times 10^{-3}$)	11.20×10^{-2} ($\pm 4.2 \times 10^{-2}$)
Nonessential (n = 439)	49.59×10^{-3} ($\pm 2.1 \times 10^{-3}$)	5.93×10^{-3} ($\pm 0.6 \times 10^{-3}$)	13.08×10^{-2} ($\pm 0.9 \times 10^{-2}$)
Undetermined (n = 599)	45.51×10^{-3} ($\pm 1.7 \times 10^{-3}$)	5.51×10^{-3} ($\pm 0.5 \times 10^{-3}$)	12.60×10^{-2} ($\pm 0.8 \times 10^{-2}$)
Significance of the difference[c] essential vs. nonessential	NS	NS	NS
Significance of the difference[c] essential vs. undetermined	NS	NS	NS
Significance of the difference[c] nonessential vs. undetermined	NS	NS	NS

[a]The average synonymous substitution (nucleotide substitutions that do not change the encoded amino acid sequence) rate (Ks) for all orthologous genes within a given fitness class and species is shown with the standard error (in parentheses).
[b]The average nonsynonymous substitution (nucleotide substitutions that change the encoded amino acid sequence) rate (Ka) for all orthologous genes within a given fitness class and species is shown with the standard error (in parentheses).
[c]Statistical significance of the difference between the essential and nonessential classes for a given measurement as determined using the Mann-Whitney *U* test.

diction by comparing the evolutionary rate of essential versus nonessential genes have yielded equivocal results. This has led to the speculation that the rate of evolution for a given gene is determined more by the proportion of amino acid residues in the encoded protein that are critical for maintaining function than by the magnitude of the selection coefficient against deleterious mutations in that gene (Brookfield 2000). However, these two factors, in reality, might not be independent. In light of the results reported here, it might be the case that at least for bacterial genes, the rate of protein evolution is determined by the proportion of sites in a protein that has a large selection coefficient against deleterious mutations.

METHODS

Classification of *E. coli* K12 genes as essential, nonessential, or undetermined was taken from the PEC database (http://www.shigen.nig.ac.jp/ecoli/pec/). The PEC database classifies genes as essential or nonessential on the basis of a combination of experimental evidence and general functional considerations. If a strain has a null mutation in a gene and is able to grow, the gene in question is considered to be nonessential. Genes for which conditional lethal mutants have been isolated (Chow and Berg 1988; Harris et al. 1992) are classified as essential. In addition to the experimentally characterized genes, a much smaller subset of *E. coli* K12 genes were classified as essential or nonessential based on their functional

characteristics (in absence of the specific experimental data listed above). For example, ribosomal structural genes and genes encoding unique aminoacyl-tRNA synthetases are classified as essential. Conversely, genes involved in flagellation, chemotaxis, and mobility were classified as nonessential. Essential, nonessential, and undetermined genes were placed into four broad functional categories using the classification scheme used in the COGs database (Tatusov et al. 1997, 2000, 2001).

Proteobacterial species with more than one complete genome sequence available as of GenBank release 123.0 were analyzed here: *E. coli* K12 (Blattner et al. 1997), *E. coli* O157:H7 (Perna et al. 2001), *H. pylori* 26695 (Tomb et al. 1997), *H. pylori* J99 (Alm et al. 1999), *N. meningitidis* serogroup B strain MC58 (Tettelin et al. 2000), and *N. meningitidis* serogroup A strain Z2491 (Parkhill et al. 2000). Nucleotide sequences and protein sequences predicted from complete bacterial genomes were obtained from the National Center for Biotechnology Information (NCBI) FTP server (ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria).

Orthologous protein sequences encoded by complete intraspecific genomes were identified using an all-against-all BLAST (Altschul et al. 1990, 1997) procedure. The SEALS package (Walker and Koonin 1997) was used to implement multiple BLAST searches and to postprocess the results of these searches. For any two intraspecific genomes, two proteins were considered orthologs if they were symmetrical best hits in each reciprocal all-against-all BLAST search. BLAST searches were run with a bits-per-position cutoff of 0.7. Orthologous proteins were aligned using ClustalW (Thompson et al. 1994) with default options. Orthologous protein-encoding nucleotide sequences were obtained from the NCBI FTP server and aligned to correspond to the protein sequence alignments using SEALS. The number of synonymous nucleotide substitutions per synonymous site (K_s), and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (K_a) were estimated for the resulting orthologous nucleotide sequence alignments using the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993).

Orthologous protein sequences encoded by interspecific genomes (*E. coli*, *H. pylori*, and *N. meningitidis*) were identified using a similar all-against-all BLAST approach implemented with the SEALS package. Proteins that formed mutually consistent triangles of symmetrical best hits (Tatusov et al. 1997) were considered to be orthologs (hence an identical number of orthologs from each genome in this analysis). Orthologous protein sequences were aligned using ClustalW as described above, and the resulting sequence alignments were used to calculate evolutionary distances between orthologous proteins. Distances were calculated using the 3BRANCH program (Y.I. Wolf, unpubl.; available on request) that implements a distance correction for multiple hits based on the γ-distribution of site rate variation (Ota and Nei 1994; Grishin 1995) with an α-parameter of 1.0. Distances are expressed as the number of substitutions per position.

The phyletic distribution of orthologous proteins was determined using the COGs database (Tatusov et al. 1997, 2000, 2001). The COGs database at the time of this work included species that fell into 26 distinct taxonomic groups. For each orthologous protein, a phyletic distribution parameter, which is equal to the number of taxonomic groups represented in the corresponding COG, was calculated.

The bootstrap analysis was performed by resampling with replacement from the original sets of per gene values of Ka, Ks, and Ka/Ks for essential and nonessential genes (calculated as described above). For each of 1000 bootstrap replicates, a resampled set with the same number of values as the original set was constructed. The average values for these resampled sets were then calculated and the averages of the essential and nonessential sets were compared.

Levels of significance for the difference among average Ks, Ka, Ka/Ks, phyletic distribution parameter, and levels of protein sequence variation were determined using the Mann-Whitney *U* test.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., et al. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397:** 176–180.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

Brookfield, J.F. 2000. What determines the rate of sequence evolution? *Curr. Biol.* **10:** R410–R411.

Chow, W.Y. and Berg, D.E. 1988. Tn5tac1, a derivative of transposon Tn5 that generates conditional mutations. *Proc. Natl. Acad. Sci.* **85:** 6468–6472.

Grishin, N.V. 1995. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* **41:** 675–679.

Harris, S.D., Cheng, J., Pugh, T.A., and Pringle, J.R. 1992. Molecular analysis of *Saccharomyces cerevisiae* chromosome, I: On the number of genes and the identification of essential genes using temperature-sensitive-lethal mutations. *J. Mol. Biol.* **225:** 53–65.

Hirsh, A.E. and Fraser, H.B. 2001. Protein dispensability and rate of evolution. *Nature* **411:** 1046–1049.

Hurst, L.D. and Smith, N.G. 1999. Do essential genes evolve slowly? *Curr. Biol.* **9:** 747–750.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, NewYork, NY.

Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.

———. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.

Li, W.H. and Graur, D. 1991. *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, MA.

Makalowski, W. and Boguski, M.S. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci.* **95:** 9407–9412.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23:** 263–286.

Ohta, T. and Ina, Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *J. Mol. Evol.* **41:** 717–720.

Ota, T. and Nei, M. 1994. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* **38:** 642–643.

Pamilo, P. and Bianchi, N.O. 1993. Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. *Mol. Biol. Evol.* **10:** 271–281.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., et al. 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404:** 502–506.

Perna, N.T., Plunkett III, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409:** 529–533.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28:** 33–36.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A.,

Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:** 22–28.
Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., et al. 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287:** 1809–1815.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.
Walker, D.R. and Koonin, E.V. 1997. SEALS: A system for easy analysis of lots of sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 333–339.
Wilson, A.C., Carlson, S.S., and White, T.J. 1977. Biochemical evolution. *Annu. Rev. Biochem.* **46:** 573–639.
Wolfe, K.H. and Sharp, P.M. 1993. Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37:** 441–456.

WEB SITE REFERNCES

ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria; Bacterial genomes from the National Center for Biotechnology Information FTP server.
http://www.shigen.nig.ac.jp/ecoli/pec/; Profiling of the *E. coli* Genome database.

Received January 23, 2002 ; accepted in revised form March 25, 2002.